DOCUMENT RESUME

ED 336 046 HE 024 868

AUTHOR Pike, Gary R.
TITLE Lies, Damn Lies, and Statistics Revisited: A Comparison of Three Methods of Representing Change. AIR 1991 Annual Forum Paper.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE May 91
NOTE 28p.; Paper presented at the Annual Forum of the Association for Institutional Research (31st, San Francisco, CA, May 26-29, 1991).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; College Seniors; Error of Measurement; Higher Education; Institutional Research; *Measurement Techniques; Research Methodology; *Statistical Analysis; Test Interpretation; *Test Reliability
IDENTIFIERS *AIR Forum; Gain Scores; Repeated Measures Design; Residual Scores; University of Tennessee Knoxville

ABSTRACT

Because change is fundamental to education and the measurement of change assesses the quality and effectiveness of postsecondary education, this study examined three methods of measuring change: (1) gain scores; (2) residual scores; and (3) repeated measures. Data for the study was obtained from transcripts of 722 graduating seniors at the University of Tennessee, Knoxville. The gain method involves administering an instrument at the beginning of a program of study and then readministering the instrument at the end of the program. Residual scores are calculated by regressing students' scores at the end of a program of study on their entering scores in order to develop a prediction model. The difference between actual and predicted scores then represents student change. The repeated measures method uses all of the data from the two tests to describe change. Results of the analysis and comparison found that all three methods were marred by similar problems of unreliability. Reliability coefficients and large standard errors of measurement suggested that what is being measured is not true change, but error. However, the repeated measures technique offered the greatest potential because it maintains the original test data, allowing researchers to bring more information to bear in interpreting their findings. One table, one figure and 41 references accompany the text. (JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

1

Lies, Damn Lies, and Statistics[1] Revisited:

A Comparison of Three Methods of Representing Change*

ED 326046

Gary R. Pike

Associate Director,

Center for Assessment Research & Development

University of Tennessee, Knoxville

1819 Andy Holt Avenue

Knoxville, Tennessee 37996-4350

(615) 974-2350

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual forum of the

Association for Institutional Research

San Francisco, May 1991

HE 024 868

* Approximately 10% of the cost of this research was supported by a $2000 grant from the Fund for the Improvement of Postsecondary Education (FIPSE) via the AAHE Assessment Forum.

BEST COPY AVAILABLE

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-First Annual Forum of the Association for Institutional Research held at The Westin St. Francis, San Francisco, California, May 26-29, 1991. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Chair and Editor
Forum Publications Editorial
Advisory Committee

Abstract

Numerous authors have argued that change is fundamental to the education process, and that the measurement of change is an essential element in efforts to assess the quality and effectiveness of postsecondary education. Despite widespread support for the concept of studying student growth and development, many researchers have been critical of existing methods of representing change. Intended for assessment practitioners and educational researchers, this study examines three methods of measuring change: (1) gain scores, (2) residual scores, and (3) repeated measures. Analyses indicate that all three methods are seriously flawed, although repeated measures offer the greatest potential for adequately representing student growth and development.

Lies, Damn Lies, and Statistics[1] Revisited: A Comparison of Three Methods of Representing Change

Numerous authors have argued that change is fundamental to the education process, and that the measurement of change is an essential element in efforts to assess the quality and effectiveness of postsecondary education (Astin, 1987; Ewell, 1984; Linn, 1981; Nuttall, 1986). Astin (1987) has also argued that change scores are superior to outcomes measures because they control for the effects of differences in students' entering characteristics. Other reported advantages of change scores include their usefulness in encouraging faculty involvement in assessment and their appropriateness for evaluating nontraditional programs and students (Astin & Ewell, 1985; McMillan, 1988; Vaughan & Templin, 1987).

Critics of change research have focused on the practical problems associated with the measurement of student growth and development. Warren (1984) has identified three specific problems: (1) in many instances, students do not have a sufficient knowledge base against which change can be measured; (2) when significant differences in the knowledge base are present, scores cannot be compared; and (3) measurement technology is not sufficiently advanced to assess the effects of education on student growth. These and other problems led Cronbach and Furby (1970, p. 78) to conclude: "There appears to be no need to use measures of change as dependent variables and no virtue in using them."

Given the diversity of opinion about the value of studying change, drawing generalizations from the literature is extremely difficult. However, one point is clear: If studies of change are to be used to evaluate education programs, researchers must carefully evaluate existing methods of representing change and identify ways in which those methods can be improved (Everson, 1986; Fincher, 1985; Pascarella, 1989).

Statisticians and educational researchers have identified several methods for measuring student growth and development. This essay examines three of those methods: (1) gain scores; (2) residual scores; and (3) repeated measures. Initially, a brief description of each method is provided.

These descriptions are followed by an analysis of freshman-to-senior gains using each method. Finally, the strengths and weaknesses of the methods are discussed.

Methods of Representing Change

Using gain scores is the most popular method of assessing student change and is assumed to be the method of choice in assessing "value-added" or "talent- development" (Astin, 1987). In fact, gain scores are reported by the College Outcome Measures Program (COMP) staff as indicators of student learning and program effectiveness (Steele, 1988). The simplicity and intuitive appeal of gain scores is undoubtedly responsible for their popularity. Calculating a gain score entails administering an instrument at the beginning of a program of study and then readministering the instrument upon completion of the program (Baird, 1988; McMillan, 1988). The difference between the two scores is a measure of student growth, and the average across students is a measure of institutional (program) effectiveness (Steele, 1988).

Because scores at entry are subtracted from exiting scores, it is essential that they represent the same construct (Baird, 1988; Thorndike, 1966). In practice, this requirement necessitates the use of the same test, or parallel forms of a test, for both administrations. Methods of testing this assumption have been proposed by both Kristof (1969) and Lord (1957).

Residual scores (sometimes referred to as residual change scores or residual gain scores) have been used in two recent studies (Pike & Phillippi, 1989; Ratcliff, 1988). Residual scores are calculated by regressing students' scores at the end of a program of study on their entering scores in order to develop a prediction model. The difference between actual and predicted scores represents student change, while the mean represents program effects (Baird, 1988; Hanson, 1988). Because regression techniques utilize a least squares method, the sum of the residual scores for the total sample is zero (Draper & Smith, 1981; Thorndike, 1978). Consequently, the use of residual scores requires the comparison of two or more groups.

Residual scores do not require the use of the same test or parallel forms of a test. However, residual scores are a form of analysis of

covariance in which pre-test scores are used as covariates for post-test scores and, as a consequence, the assumptions underlying the analysis of covariance should be met. Two important assumptions of the analysis of covariance are that the covariates (initial scores) are unrelated to grouping variables and that the covariates are measured without error (Elashoff, 1969). Violation of these assumptions is usually seen in heterogeneity of regression slopes and may produce models for adjusting group means that overestimate group differences (Kennedy & Bush, 1985; Winer, 1971).

Unlike gain scores and residual scores, analyses of student growth and development that are based on repeated measures do not reduce change to a single score. Instead, all of the data from the two test administrations are used to describe change (Roskam, 1976). As a consequence, the original metric of the test scores can be preserved, and researchers do not have to be concerned about the accuracy of a prediction model. In addition, a variety of data analysis techniques can be applied to the data (Kennedy & Bush, 1985). Use of repeated measures does assume that the same construct is being measured over time.

A Research Example

Methods

Data from 722 graduating seniors at the University of Tennessee, Knoxville (UTK) were used to compare gain scores, residual scores, and repeated measures. Approximately 95% of the students in the research were white and 44% were male. The students were, on average, 18.2 years old when they were tested as freshmen and 22.0 years old when they were tested as seniors. The estimated Enhanced ACT Assessment composite score mean for the group was 23.1, while their mean cumulative grade point average was 3.02.

Initially, transcripts were analyzed to determine the total number of credit hours each student had completed in 90 different disciplines. Disciplines were grouped into eight categories (agriculture, business, communication, education, engineering and mathematics, humanities, natural science, and social science), and the total number of credit hours a student had completed in each category was calculated. Using scores for the eight

coursework categories, students were clustered into five groups using the within-cluster average linkage method (Aldenderfer & Blashfield, 1984). Based on an analysis of cluster means, the groups were labeled: Business and General Education (N=368), Engineering and Mathematics (N=84), Social Science and Humanities (N=180), Natural Science (N=72), and Humanities (N=18). The Humanities coursework cluster was dropped from subsequent analyses because of its small size. A discriminant analysis was able to correctly classify 95% of the students in the four remaining coursework clusters.

Data on cluster membership was merged with students' freshman and senior total scores on the College Outcome Measures Program (COMP) Objective Test. Developed as a measure of effective adult functioning, the Objective Test contains 60 questions, each with two correct answers (Forrest & Steele, 1982). Questions on the Objective Test are divided among 15 separate activities drawing on material (stimuli) from television programs, radio broadcasts, and print media. Scoring procedures for the test produce a maximum possible total score of 240 points and a chance score of 120 points (Forrest & Steele, 1982).

Students' gain scores on the Objective Test were calculated by subtracting their freshman total scores from their senior total scores. Residual scores were calculated by regressing senior total scores on freshman total scores, and results indicated that freshman scores were a strong predictor of senior scores ($F=596.99$; $df=1,702$; $p<.001$; $R^2=.46$). Using the regression coefficients from the final model, the following equation was developed to calculate residual scores: **Residual = Sr. COMP - (86.26 + (0.59 * Fr. COMP))**.

Gain scores, residual scores, and repeated measures (the original freshman and senior Objective Test scores) were analyzed using analysis of variance procedures (ANOVA) with the four coursework clusters as the classification variable. Results were interpreted as the effects of differential patterns of coursework on student growth and development.

Results

Analysis of variance results indicated that gain scores for the four coursework clusters were significantly different ($F=4.22$; $df=3,700$; $p<.01$). In contrast residual scores did not differ significantly by coursework cluster ($F=0.03$; $df=3,700$; $p>.90$). The repeated measures analysis produced results that were identical to the results for gain scores. Table 1 presents the gain score, residual score, and repeated measures (freshman and senior total score) means for the four coursework clusters and the total sample. An examination of the gain score means reveals that the gains for the Business and General Education and the Social Science and Humanities coursework clusters (13.9 and 12.5 respectively) were substantially greater than gains for either Engineering and Mathematics (10.4) or Natural Science (9.2). In contrast, small positive residual score means were observed for the Business and General Education and the Engineering and Mathematics clusters (0.1 and 0.1 respectively), while small negative means were found for Social Science and Humanities (-0.1) and Natural Science (-0.3).

Insert Table 1 about here

An examination of the freshman and senior total score means for the repeated measures analysis reveals that the initial performance of students in the Business and General Education and the Social Science and Humanities clusters (175.9 and 178.7) was much lower than the initial performance of students in the Engineering and Mathematics (184.3) and Natural Science (196.3) clusters. Differences in student performance as seniors were much smaller, although means for the Business and General Education and the Social Science and Humanities clusters (189.8 and 191.2 respectively) were still below the means for the Engineering and Mathematics and the Natural Science clusters (194.7 and 195.5 respectively). Overall, the repeated measures analyses suggested that the growth trajectories for the Business and General Education and the Social Science and Humanities coursework clusters were

steeper (indicating more growth) than the growth trajectories for Engineering and Mathematics and Natural Science.

Strengths and Weaknesses of the Three Methods

It is significant that the findings of the present research parallel results reported by Lord (1967, 1969) in which gain scores produced significant indicators of change, while residual scores (analysis of covariance) produced nonsignificant results. Given the findings for gain scores, residual scores, and repeated measures, it is useful to consider the strengths and weaknesses of the three approaches.

Gain Scores

Despite their intuitive appeal, many researchers argue that gain scores should be used with caution, if at all. One of the major problems with reliance on gain scores is that they tend to be unreliable. When a pre-test score is subtracted from a post-test score, common (reliable) variance is removed (Baird, 1988). The result is that unique variance, which includes error variance, contributes a larger share to gain score variance than to either pre-test or post-test variance. Thus, gain scores actually compound the unreliability of pre- and post-test measures.

Bereiter (1963) observed that there is an inverse relationship between the correlation of pre- and post-test scores and the reliability of difference scores. Specifically, the higher the correlation between pre- and post-test measures, the lower the reliability of their difference scores. This relationship can be clearly seen in Lord's (1963) formula for the reliability of gain scores: $r_{dd} = (r_{xx}s_x^2 + r_{yy}s_y^2 - 2r_{xy}s_xs_y) / (s_x^2 + s_y^2 - 2r_{xy}s_xs_y)$. In this formula, $r_{xx}s_x^2$ represents the true-score variance for the pre-test, while $r_{yy}s_y^2$ represents the true-score variance for the post-test. The term $2r_{xy}s_xs_y$ represents the combined variance that is common to both tests. Thus, the numerator in Lord's formula represents combined true-score variance less common variance, and the denominator represents combined total variance less common variance. Obviously, as common variance increases, the amount of true-score variance that is not common to both tests decreases as a proportion of total score variance, as does the reliability of the gain score. For the

special case where the reliability estimates and variances for the two tests are equal (i.e., the tests are parallel forms), the reliability of the gain score will approach zero as the correlation between the two tests approaches their reliability coefficient (Linn, 1981).

Several authors have suggested that there are circumstances under which a difference score will be reliable. Willett (1988) and Zimmerman and Williams (1982) argue that if there are substantial differences in the reliability estimates or standard deviations for the two tests, or if the correlation between test scores is low, the reliability of gain scores can be quite high. However, it seems appropriate to ask whether the same test, or parallel forms of the test, will produce significantly different reliability estimates or standard deviations for the two administrations. If such differences are present, or if the correlation between scores is low, it may be appropriate to ask whether the same construct is being measured at both points in time (Bereiter, 1963; Hanson, 1988; Linn & Slinde, 1977). It is not surprising that Zimmerman and Williams conclude their discussion with the caveat that their article is intended to <u>suggest</u> that gain scores <u>may</u> be reliable, not to indicate that the are reliable in practice.

Data from the present study confirm that individual gain scores are unreliable. Alpha reliability estimates for the two administrations of the COMP Objective Test are identical (.72). Standard deviations for freshman and senior total scores are 16.05 and 13.92 respectively, and the correlation between scores is .68. These values produce a reliability coefficient of .14 for individual gain scores. Given that the standard deviation of gain is 12.2, the standard error of measurement is 11.3 for individual gain scores and the 95% confidence interval for gain scores is ±22.1 points. Consequently, an observed gain score of 10 points for an individual represents a true score gain of from -12.1 to 32.1 points 95% of the time.

Linn (1981) has concluded that the poor reliability of gain scores for individuals precludes their use in decision-making, but that the problem of unreliability is less severe when a group mean is the unit of analysis. Linn's claim can be tested by establishing confidence intervals about mean

COMP scores for freshmen and seniors and then using those confidence intervals to assess the dependability of a mean gain score.

Assuming a sample size of 704 students and using the G-study variance components reported by Steele (1989), the 95% confidence interval about a true score mean is approximately ±6.2 points. (Calculation of this confidence interval is based on the fact that scores from different forms of the COMP exam are included in the group mean.) If the true score mean is 178.7 for freshmen and 191.3 for seniors, the 95% confidence intervals are 172.5 to 184.9 and 185.1 to 197.5 respectively. Consequently, observed mean gain scores will range from 0.2 (185.1 - 184.9) to 25 (197.5 - 172.5) for a sample of 704 students with a true mean gain score of 12.6 points.

A second problem with the use of gain scores is the presence of a spurious negative correlation between gain and initial status (Bereiter, 1963; Hanson, 1988; Linn, 1981; Linn & Slinde, 1977; Lord, 1963; Terenzini, 1989; Thorndike, 1966). Here again, the problem is the result of measurement error. In a gain score, the error component for pre-test scores is present, but with a negative sign due to subtraction. The presence of the pre-test error component, with opposite signs, in both the pre-test and gain scores produces a negative correlation between gain and initial status. Because the correlation is the product of measurement error, it must be considered spurious.

The data on freshman-to-senior gains in the present study provide evidence of the negative correlation between gain and initial status. The correlation between freshman scores on the Objective Test and gain scores is negative and significant (-.54). Even when mean scores for 10 undergraduate colleges are used, the negative correlation between gain and initial status is substantial (-.31). This latter result is consistent with a correlation of -.34 reported by the COMP staff (Steele, 1988).

The negative correlation between gain and initial status is particularly troublesome when the objective of assessment is to evaluate the effectiveness of education programs. The negative correlation between gain and initial status results in a built-in bias favoring individuals or programs

that perform poorly on the pre-test. If educational experiences are related to initial status, the negative correlation between gain and pre-test scores will produce relationships that are statistical artifacts, rather than true relationships (Thorndike, 1966).

In the research example, students in the Business and General Education and the Social Science and Humanities coursework clusters had the lowest scores on the Objective Test as freshmen. They also had the highest gain scores. This is not to say that coursework in business, the humanities and the social sciences does not produce significant gains. However, the negative correlation between gain and initial status makes it impossible to determine how much of the gain is due to coursework and how much is due to a statistical artifact.

Although the problems of poor reliability and a negative correlation between gain and initial status are the most frequently-cited limitations of gain scores, problems of interpretability also confound studies of gain. Both Lord (1956) and Thorndike (1966) have noted that comparisons of individuals and groups assume that the same gain scores at different parts of the score scale are numerically equivalent (i.e., a gain score of 10 points represents the same amount of change irrespective of initial status.)

Banta, Lambert, Pike, Schmidhammer, and Schneider (1987) have pointed out that students of lower initial ability tend to give incorrect responses to easier questions than do students of higher ability. (This is inherent in the definitions of item difficulty and discrimination.) Students with lower initial scores can improve by correctly answering easier items than can students with higher pre-test scores. Thus, identical gain scores may be <u>qualitatively</u> different depending on initial status. Lord (1958, p. 450) has observed: "unless the two students started at the same point in the score scales, however, it cannot be concluded that the first student really learned more than the second, except in some very arbitrary sense."

Reliance on group means does not solve problems created by the fact that gain is not numerically equivalent across the score scale. A gain score mean is a parametric statistic that assumes numerical equivalence. Since this

assumption is violated, it can be argued that mean gain scores do not meet the definitional requirements of a parametric statistic and should not be used in studies of growth and development.

Residual Scores

Because residual scores avoid problems created by the spurious negative correlation between growth and initial status, they have been used in several studies of student growth and development (Pike & Phillippi, 1989; Ratcliff, 1988). The fact that residuals are not related to pre-test results can be seen in the means from the research example. The Engineering and Mathematics coursework cluster had on of the highest pre-test means and a positive residual mean score, while the Social Science and Humanities cluster had a low pre-test mean and a negative residual mean. Students in the Business and General Education cluster had a low pre-test mean and a positive residual mean, while students in the Natural Science cluster had a high pre-test mean and a negative residual mean.

It is important to note that residuals achieve their independence from pre-test performance by creating scores that are not measures of change per se (Baird, 1988; Cronbach & Furby, 1970). Residual scores are merely that part of a post-test scores that is not linearly predictable from a pre-test score (Linn & Slinde, 1977). Furthermore, residual scores cannot be assumed to be a corrected measure of gain because the removal of pre-test effects undoubtedly eliminates some portion of true change (Cronbach & Furby, 1970).

One thing that residual scores clearly do not accomplish is to significantly improve the reliability with which change is measured (Linn & Slinde, 1977). Based on his review, Willett (1988) concludes that the reliability of residual scores is not much different from the reliability of gain scores. Using data from the research example and Traub's (1967) preferred method of calculating the reliability of residual scores produces a coefficient of .17, which is not much different from the reliability coefficient of .14 for gain scores.

Residual scores also suffer from problems of interpretability because they represent deviations from an average prediction for the total sample.

When residuals are used, approximately half of the students will be above average and half will be below average. If one program is judged to be effective, another will be judged to be ineffective, even though both may be doing an exemplary job (Baird, 1988). Compounding this problem is the fact that the prediction model may be inaccurate. Baird (1988) has noted that even correlations of .60 between pre- and post-tests can yield highly inaccurate prediction models.

A practical issue related to the interpretation of residuals is that of what variables to include in the prediction model. Including students' background characteristics along with their pre-test scores can improve the accuracy of the prediction model, but it further removes residual scores from the realm of a change measure (Baird, 1988). To date, no criteria are available to guide researchers in the inclusion of variables in the regression equation.

Repeated Measures

Given the problems encountered when trying to adjust pre-test scores to control for group differences, a superior alternative might be to accept the fact that differences in initial status exist and to proceed from that point. Repeated measures analyses, however, cannot avoid the requirement that the same construct be measured over time, nor can they overcome problems created by the fact that change is numerically equivalent across the score range. Likewise, repeated measures do not eliminate situations in which individuals or groups with lower pre-test scores gain more than individuals or groups with higher pre-test scores. In repeated measures analyses, the negative correlation between gain and initial status is manifest in steeper growth trajectories for groups with lower pre-test scores.

The fact that repeated measures maintain all of the data about test performance does add a dimension of interpretation that is not available with gain scores. As previously observed, students in the Business and General Education coursework cluster gained substantially more than students in the Natural Science cluster. However, the post-test performance of students in

the Business and General Education cluster was still well below the post-test performance of students in the Natural Science cluster.

One aspect of repeated measures that is seldom discussed is the fact that they face the same problems of unreliability as gain scores. Figure 1 displays the pre- and post-test means for the total sample in the research example. These means are mid-points in a range representing the 95% confidence intervals for pre- and post-test scores. The solid line represents the growth trajectory for observed means, while the broken lines represent the range of possible growth trajectories described by the confidence intervals.

Insert Figure 1 about here

From the data in Figure 1, it is easy to see that a variety of growth trajectories are possible by chance alone. The line from the top of the pre-test range (184.9) to the bottom of the post-test range (185.1) is nearly flat and is identical to the lower confidence interval for mean gain scores (0.2 points). Conversely, the line from the bottom of the pre-test range (172.5) to the top of the post-test range (197.5) is identical to the upper confidence interval for mean gain scores (25 points). The problem of unreliability helps explain why ANOVA results for gain scores and repeated measures are identical.

Conclusions About the Three Methods

Despite the disparate results produced by the three measures of change reviewed in this paper, gain scores, residual scores, and repeated measures all face similar problems of unreliability. Reliability coefficients ranging from .14 to .17, coupled with large standard errors of measurement for both individuals and groups, suggest that what is being measured is not true change, but error. Consequently, institutional researchers should exercise caution in making judgments about the performance of either individuals or groups using measures of change.

One of the often-cited advantages of measures of change is that they control for the effects of differences in pre-test characteristics,

particularly ability. The present study clearly indicates that differences in initial status are not adequately accounted for by either gain scores or repeated measures. These approaches under-correct for initial differences and produce a spurious negative relationship between pre-test scores and growth. In contrast, residual scores over-correct for differences in initial status, eliminating both spurious pre-test effects and true gain that is linearly related to pre-test scores. The fact that gain scores and repeated measures under-correct for differences in initial status, while residual scores over-correct for these differences, helps explain why gain scores and repeated measures produce statistically significant results, while residual scores produce nonsignificant results (Bereiter, 1963). Unfortunately, it is impossible to determine which approach provides the most accurate assessment of change because "there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled pre-existing differences between groups" (Lord, 1967, p. 305).

In interpreting the effects of initial ability, repeated measures are generally superior to the other methods of representing change because they maintain the original test data, allowing researchers to bring more information to bear in interpreting their findings. However, repeated measures, along with gain scores and residuals, face other problems of interpretability. Comparisons based on either gain scores or repeated measures may be confounded by the fact that gains are qualitatively different at different points on the score continuum. Residuals, because they are deviation scores, may inform researchers that two groups of students are different, but they do not indicate whether student performance is satisfactory or unsatisfactory (Baird, 1988).

In summary, gain scores, residual scores, and repeated measures all pose serious statistical problems that argue against their use in evaluating either students or programs. Of the three methods, repeated measures are somewhat perferable because they utilize all of the data on test performance to inform decisions. If error could be eliminated from repeated measures, the measurement of change would be enhanced because problems of unreliability and

the spurious relationship between growth and initial status could be eliminated. However, even repeated measures of true scores would not overcome problems created by qualitative differences in growth at different levels of initial ability.

Certainly the findings presented in this paper are not encouraging. Although the claim of Cronbach and Furby (1970) that growth should not be a variable in institutional research is probably overstated, Warren's (1984) conclusion that measurement technology is not sufficient to accurately represent growth and development is supported by this study. Institutional researchers interested in studying student growth and development need to consider alternatives to the three traditional measures of change if they are to provide meaningful evaluations of students and/or education programs.

Notes

[1] Part of the title for this paper is borrowed from Michael Wheeler. (1976). Lies, Damn Lies, and Statistics: The Manipulation of Public Opinion in America. New York: Dell Publishers.

References

Aldenderfer, Mark S., and Blashfield, Roger K. (1984). Cluster Analysis (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-044). Beverly Hills, CA: Sage.

Astin, Alexander W. (1987). Achieving Educational Excellence. San Francisco: Jossey-Bass.

Astin, Alexander W., and Ewell, Peter T. (1985). The value added debate ... continued. AAHE Bulletin 37(8): 11-13.

Baird, Leonard L. (1988). Value-added: Using student gains as yardsticks of learning. In Clifford Adelman (ed.), Performance and Judgment: Essays on Principles and Practice in the Assessment of College Student Learning (pp. 205-216). Washington, D.C.: U.S. Government Printing Office.

Banta, Trudy W., Lambert, E. Warren, Pike, Gary R., Schmidhammer, James L., and Schneider, Janet A. (1987). Estimated student score gain on the ACT COMP exam: Valid tool for institutional assessment? Research in Higher Education 27(3): 195-217.

Bereiter, Carl (1963). Some persisting dilemmas in the measurement of change. In Chester W. Harris (ed.), Problems in Measuring Change (pp. 3-20). Madison, WI: University of Wisconsin Press.

Cronbach, Lee J., and Furby, Lita (1970) How should be measure "change" - or should we? Psychological Bulletin 74(1): 68-80.

Draper, Norman R., and Smith, Harry (1981). Applied Regression Analysis (2d ed.). New York: John Wiley.

Elashoff, Janet D. (1969). Analysis of covariance: A delicate instrument. American Educational Research Journal 6(3): 383-401.

Everson, Howard T. (1986). Where is the value in "value-added" testing? In Kathleen McGuiness (ed.), Legislative Action and Assessment: Reason and Reality (pp. 157-166). Fairfax, VA: George Mason University.

Ewell, Peter T. (1984). The Self-Regarding Institution: Information for Excellence. Boulder, CO: National Center for Higher Education Management Systems.

Fincher, Cameron (1985). What is value-added education? Research in Higher Education 22(4): 395-398.

Forrest, Aubrey, and Steele, Joe M. (1981). Defining and Measuring General Education Knowledge and Skills. Iowa City, IA: American College testing Program.

Hanson, Gary R. (1988). Critical issues in the assessment of value added in education. In Trudy W. Banta (ed.), Implementing Outcomes Assessment: Promise and Perils (New Directions for Institutional Research, series no. 59, pp. 56-68). San Francisco: Jossey-Bass.

Kennedy, John J., and Bush, Andrew J. (1985). An Introduction to the Design and Analysis of Experiments in Behavioral Research. New York: University Press of America.

Kristof, Walter (1969). Estimation of true score and error variance for tests under various equivalence assumptions. Psychometrika 34(4): 489-507.

Linn, Robert L. (1981). Measuring pretest-posttest performance changes. In Ronald A. Berk (ed.), Educational Evaluation Methodology: The State of the Art (pp. 84-109). Baltimore: Johns Hopkins University Press.

Linn, Robert L., and Slinde, Jeffrey A. (1977). The determination of the significance of change between pre- and posttesting periods. Review of Educational Research (47(1): 121-150.

Lord, Frederic M. (1956). The measurement of growth. Educational and Psychological Measurement 16(3): 421-437.

Lord, Frederic M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika 22(3): 207-220.

Lord, Frederic M. (1958). Further problems in the measurement of growth. Educational and Psychological Measurement 18(3): 437-451.

Lord, Frederic M. (1963). Elementary models for measuring change. In Chester W. Harris (ed.), Problems in Measuring Change (pp. 21-38). Madison, WI: University of Wisconsin Press.

Lord, Frederic M. (1967). A paradox in the interpretation of group comparisons. Psychological Bulletin 68(5): 304-305.

Lord, Frederic M. (1969). Statistical adjustments when comparing preexisting groups. Psychological Bulletin 72(5): 336-337.

McMillan, James H. (1988). Beyond value-added education: Improvement alone is not enough. Journal of Higher Education 59(5): 564-579.

Nuttall, Desmond L. (1986). Problems in the measurement of change. In Desmond L. Nuttall (ed.), Assessing Educational Achievement (pp. 153-167). London: Falmer Press.

Pascarella, Ernest T. (1989). Methodological issues in assessing the outcomes of college. In Cameron Fincher (ed.), Assessing Institutional Effectiveness: Issues, Methods, and Management (pp. 19-32). Athens, GA: University of Georgia Press.

Pike, Gary R., and Phillippi, Raymond H. (1989). Generalizability of the differential coursework methodology: Relationships between self-reported coursework and performance on the ACT-COMP exam. Research in Higher Education 30(3): 245-260.

Ratcliff, James L. (1988). Developing a cluster-analytic model for identifying coursework patterns associated with general learned abilities of college students. Paper presented at the annual meeting of the American Educational Research Association, New Orleans

Roskam, Edward E. (1976). Multivariate analysis of change and growth: Critical review and perspectives. In Dato N. M. De Gruijter and Leo J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement (pp. 111-133). New York: John Wiley.

Steele, Joe M. (1988). Using measures of student outcomes and growth to improve college programs. Paper presented at the annual forum of the Association for Institutional Research, Phoenix.

Steele, Joe M. (1989). College Outcome Measures Program (COMP): A generalizability analysis of the COMP Objective Test (Form 9). Unpublished manuscript, American College Testing Program, Iowa City, IA.

Terenzini, Patrick T. (1989). Measuring the value of college: Prospects and problems. In Cameron Fincher (ed.), Assessing Institutional Effectiveness: Issues, Methods, and Management (pp. 33-47). Athens, GA: University of Georgia Press.

Thorndike, Robert L. (1966). Intellectual status and intellectual growth. Journal of Educational Psychology 57(3): 121-127.

Thorndike, Robert M. (1978). Correlational Procedures for Research. New York: Gardner.

Traub, Ross E. (1967). A note on the reliability of residual change scores. Journal of Educational Measurement 4(4): 253-256.

Vaughan, George B., and Templin, Robert G., Jr. (1987). Value added: Measuring the community college's effectiveness. Review of Higher Education 10(3): 235-245.

Warren, Jonathan (1984). The blind alley of value added. AAHE Bulletin 37(1): 10-13.

Willett, John B. (1988). Questions and answers in the measurement of change. In Ernst Z. Rothkopf (ed.), Review of Research in Education (vol 15, pp. 345-422). Washington, D.C.: American Educational Research Association.

Winer, B. J. (1971). Statistical Principles in Experimental Design (2d ed.). New York: McGraw-Hill.

Zimmerman, Donald W., and Williams, Richard H. (1982). Gain scores can be highly reliable. Journal of Educational Measurement 19(2): 149-154.

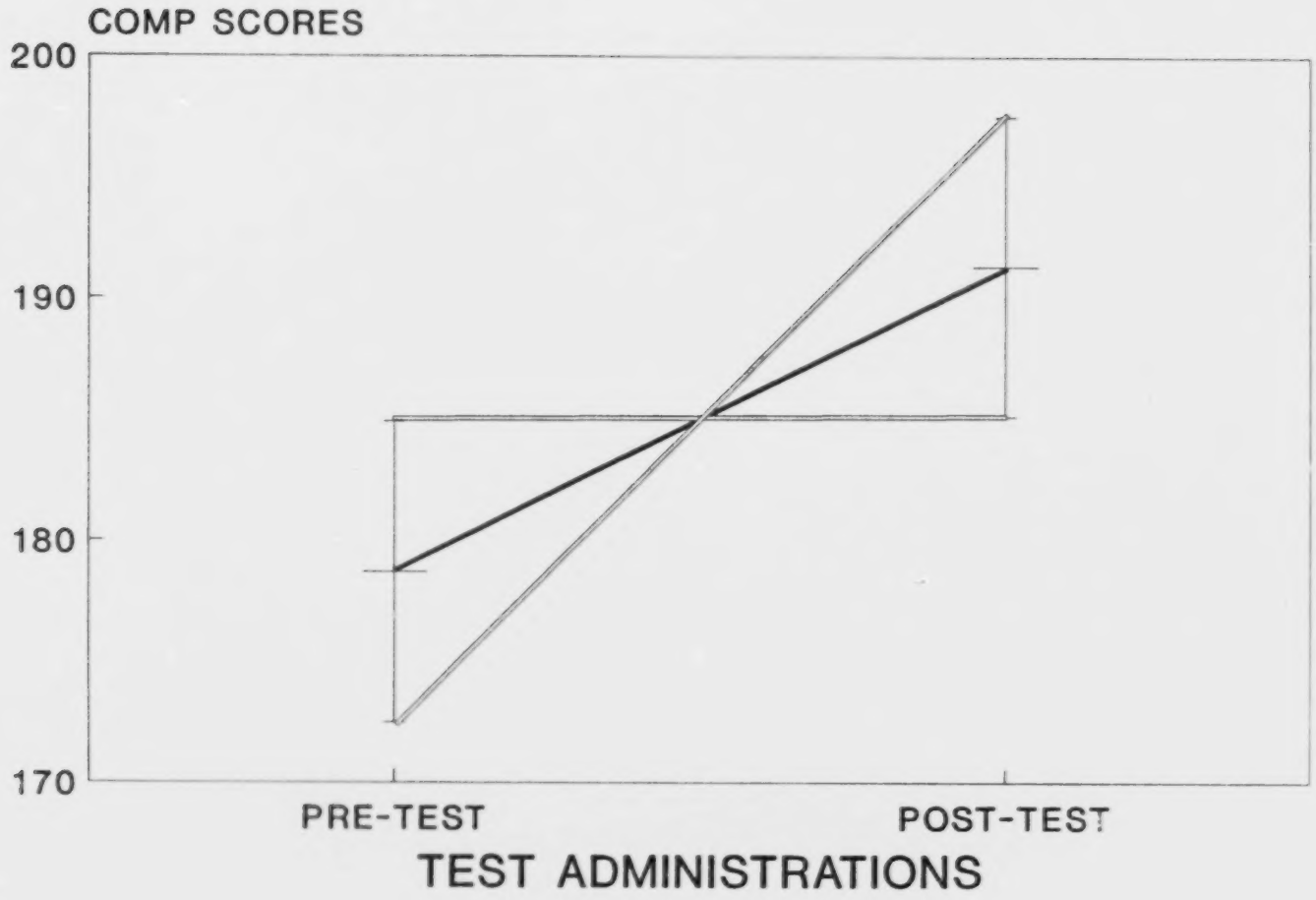
Table 1

Means for Gain Scores, Residual Scores, and Repeated Measures by Coursework Cluster

Coursework Cluster	Gain Score	Residual	Pre-Test	Post-Test
Business & Gen. Educ.	13.9	0.1	175.9	189.8
Engineering & Math.	10.4	0.1	184.3	194.7
Social Sci. & Humanities	12.5	-0.1	178.7	191.2
Natural Science	9.2	-0.3	186.3	195.5
TOTAL SAMPLE	12.6	0.0	178.7	191.3

Figure Captions

Figure 1: Possible Growth Trajectories for Freshman-to-Senior Gains


COMP SCORES
200
190
180
170
PRE-TEST
POST-TEST
TEST ADMINISTRATIONS